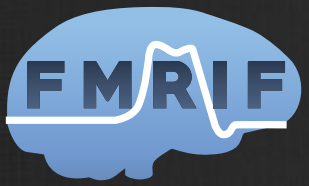
Open and Reproducible Science Using Python and GitHub

Adam Thomas & John Lee
Data Science and Sharing Team, FMRIF, NIMH



Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

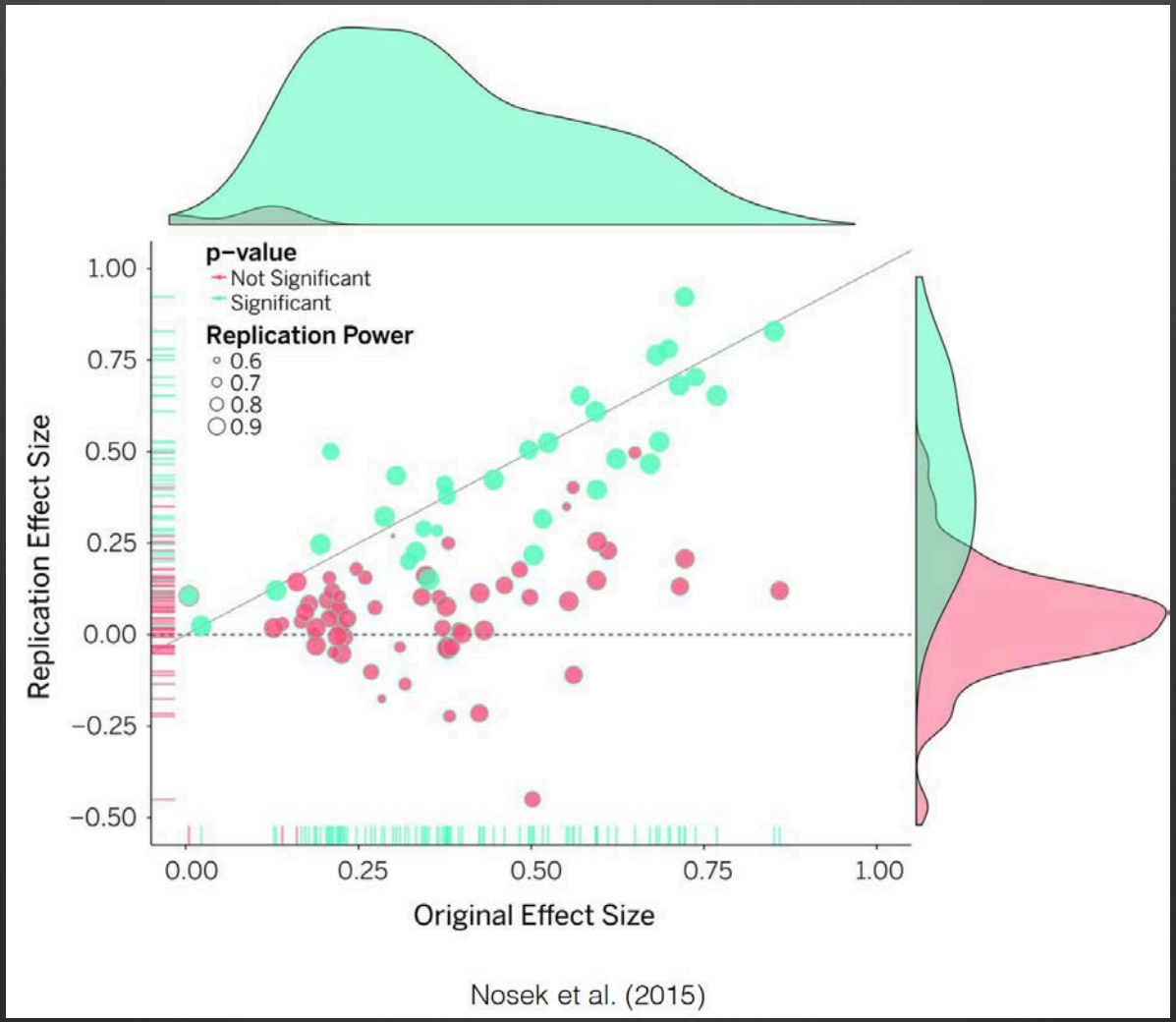
Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

Outline

- What Problems are we trying to solve?
- What is Open Science?
- How do I do Open Science?

Problem: Reproducibility



Nosek et al. (2015)

The Problem: Reproducibility

PLoS Medicine | www.plosmedicine.org

August 2005 | Volume 2 | Issue 8 | e124

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

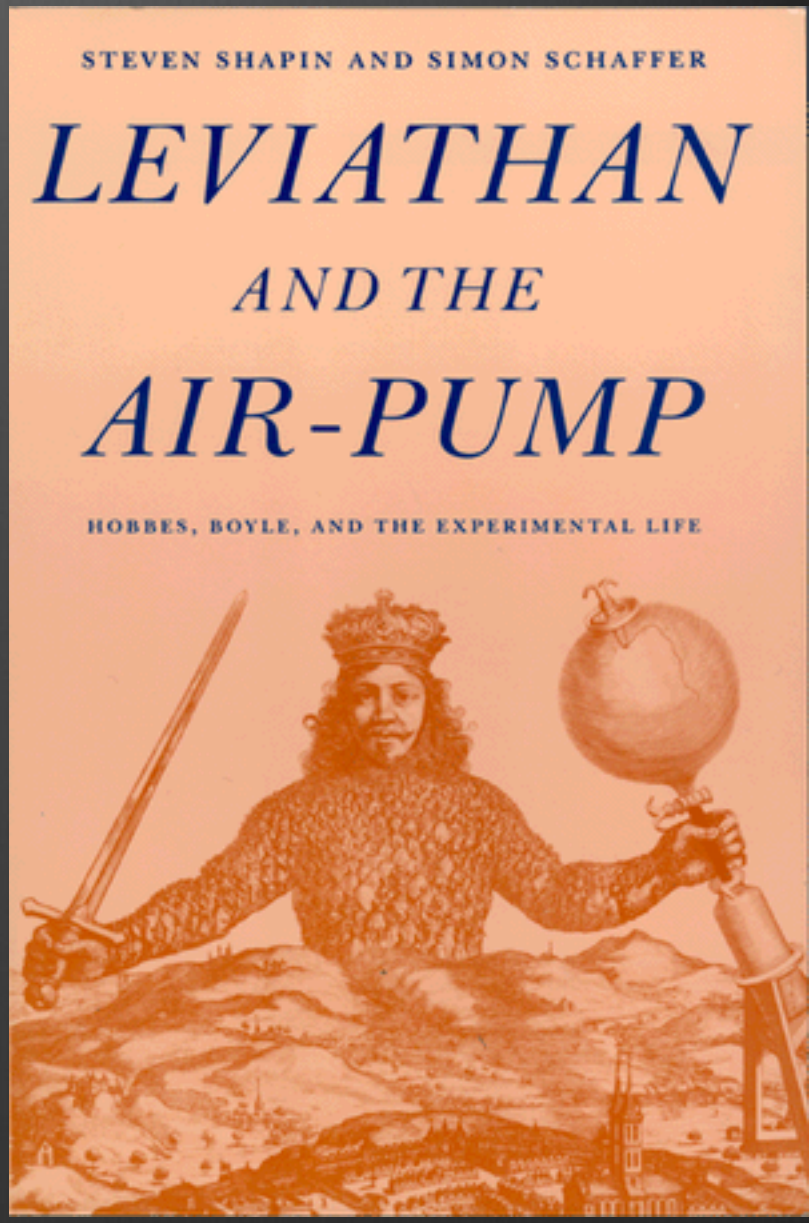
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research

Why: Replicable vs. Reproducible

- Replication is the foundation of science
- "By repeating the same experiment over and over again, the certainty of fact will emerge."
  – Robert Boyle



Problem:
Wasted time & resources

"How much time do you spend handling, reorganizing, and managing your data as opposed to actually *doing* science?"

- Median answer is 80%

Problem:
Wasted Time & resources

Unpublished Data

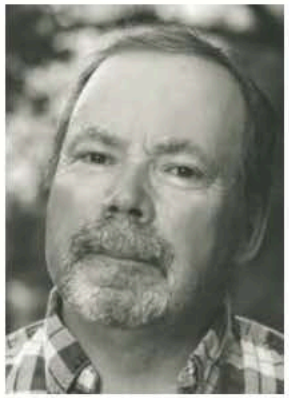
- File drawer problem
- Lost staff & lost metadata
- Underutilized data

Problems: The big-data revolution

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne, Jon R. Lorsch** *and* **Eric D. Green.**



Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-

recorded. All of this means that absolute numbers are hard to interpret.

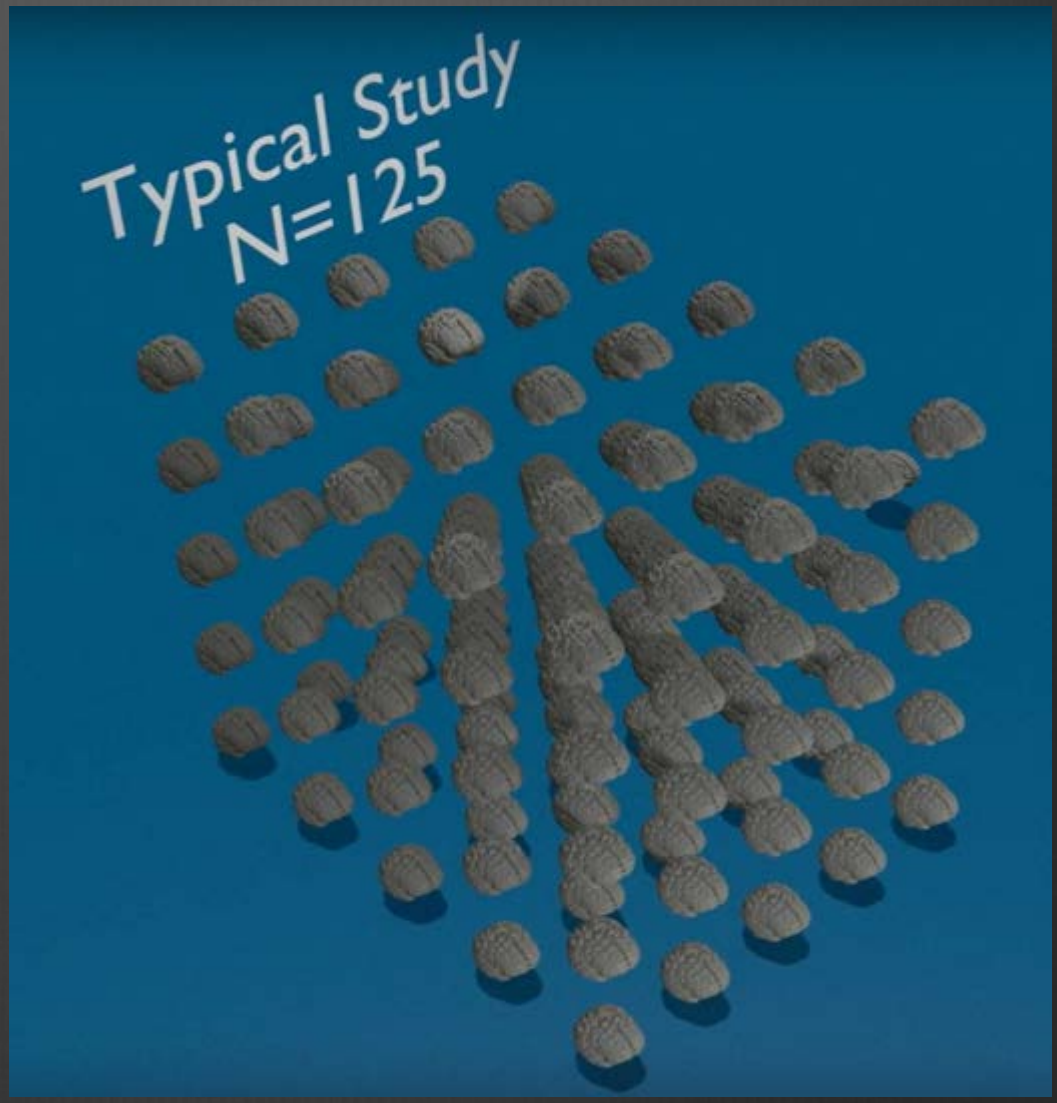
These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

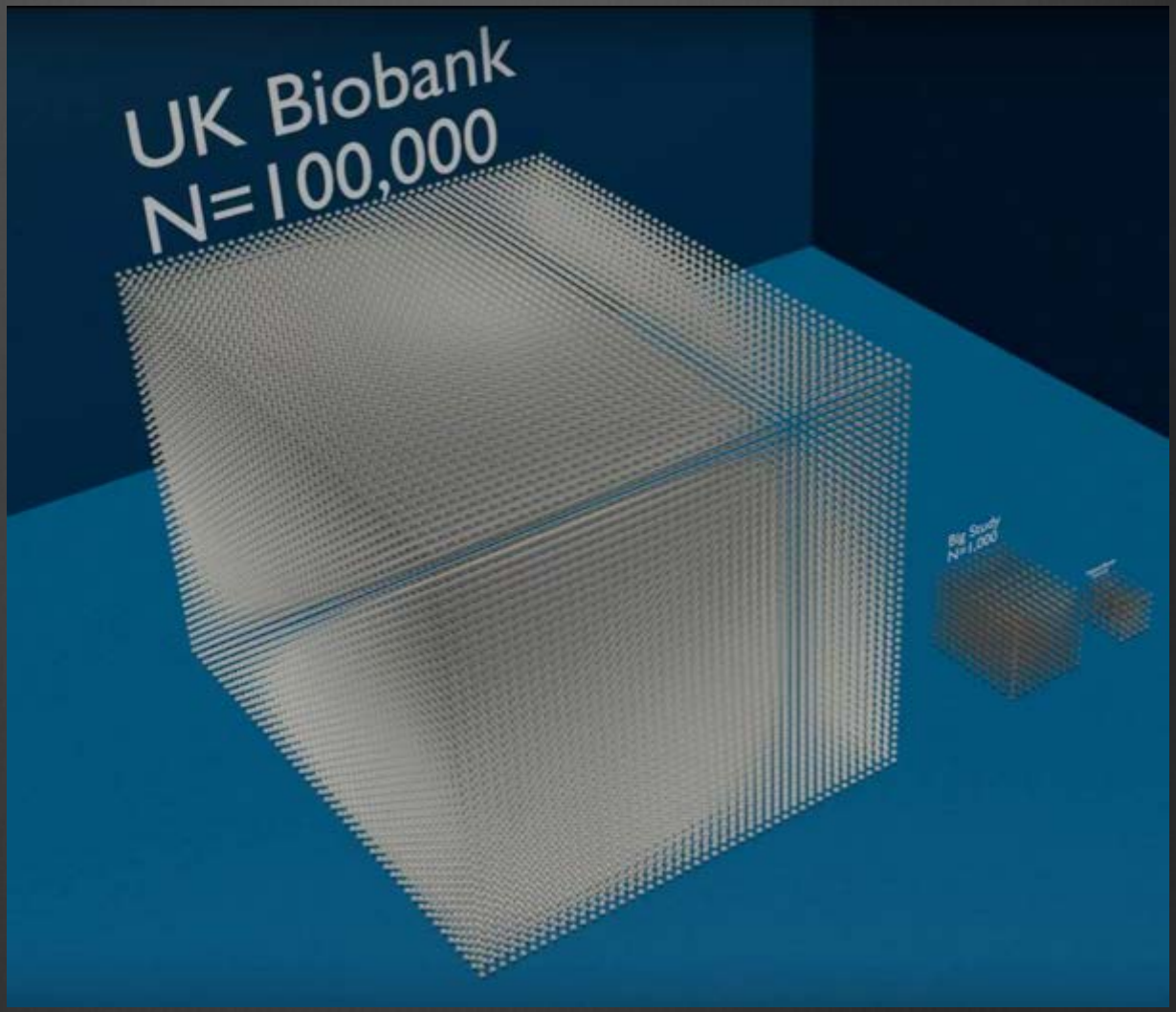
Problems: The big-data revolution

UK
Biobank
Imaging
Initiative



Problems: The big-data revolution

UK Biobank Imaging Initiative



Problems: The big-data revolution

Obama's precision medicine initiative will aim to enroll a large number of people in a genetic database representing the U.S. population.

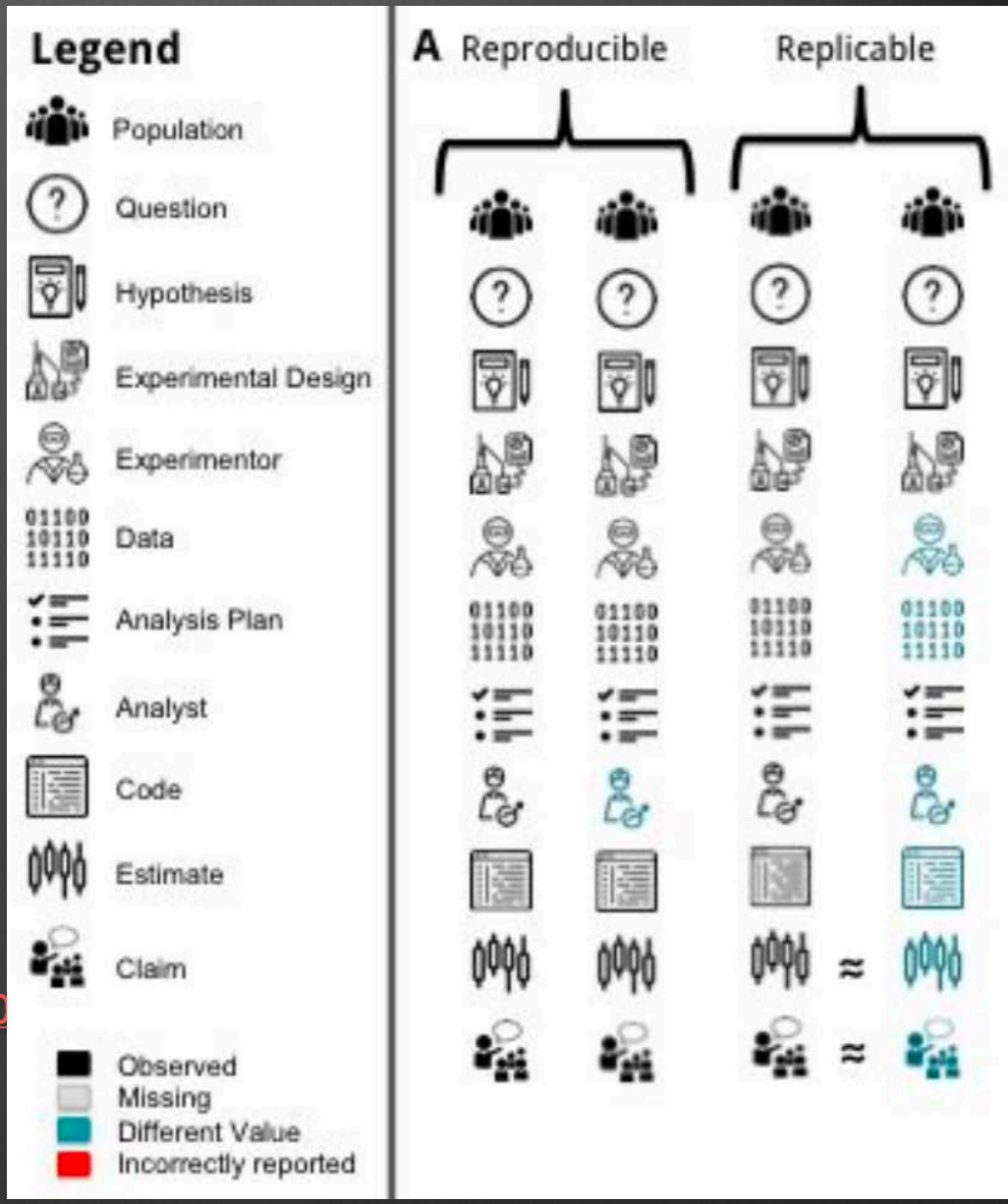
Amy West/Flickr (CC BY 2.0)

President Obama's 1-million-person health study kicks off with five recruitment centers

By Jocelyn Kaiser | Jul. 7, 2016 , 5:00 PM

Why: Replicable vs. Reproducible

- Replication: new data, different team
- Reproducible: Same data, different team
- Where replication is hard or impossible, reproducible is the next best thing
- http://biorxiv.org/content/biorxiv/early/2016/07/29/066803.full.pdf



Outline

- The Problem
  - Reproducibility
  - Wasted resources
  - Lack of integration
  - Ill-prepared to work with big datasets
- What is Open Science?
- How do I do Open Science?

Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

What is Open Science?

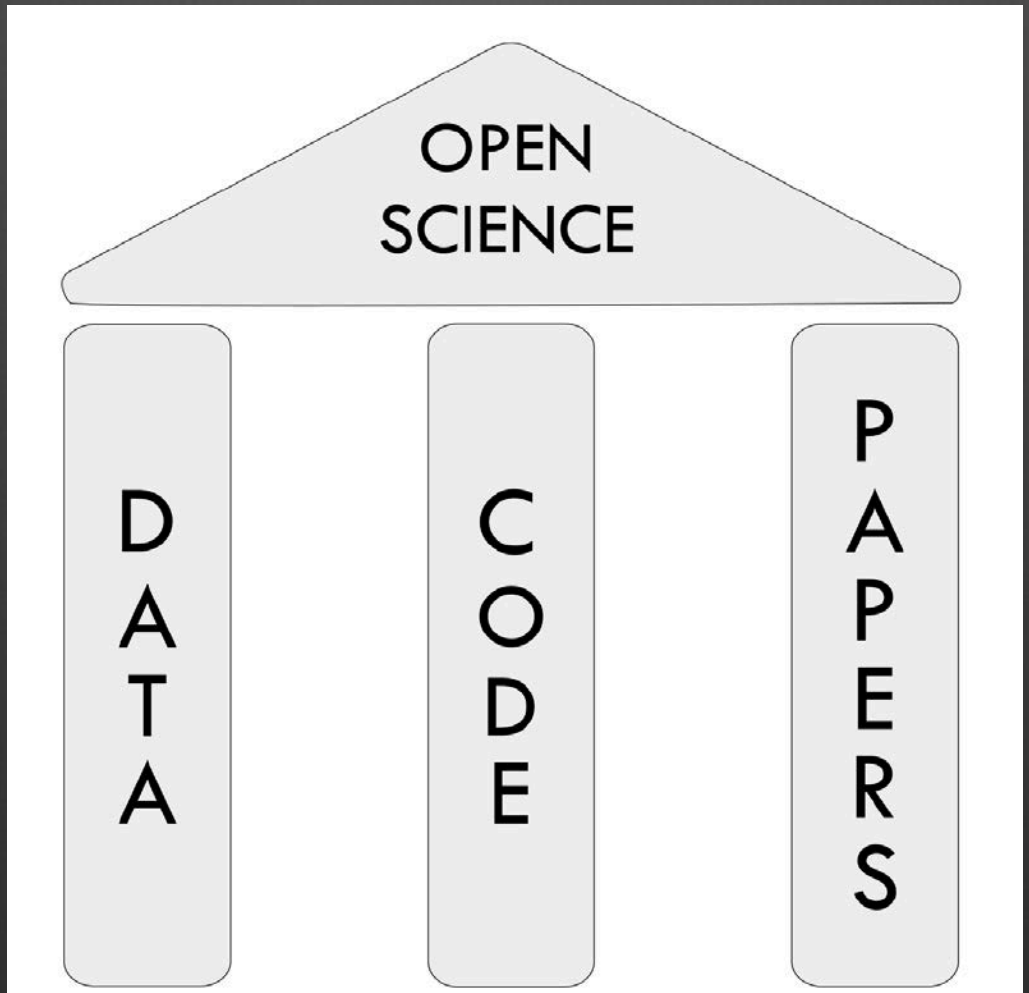


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

What is Open Data?

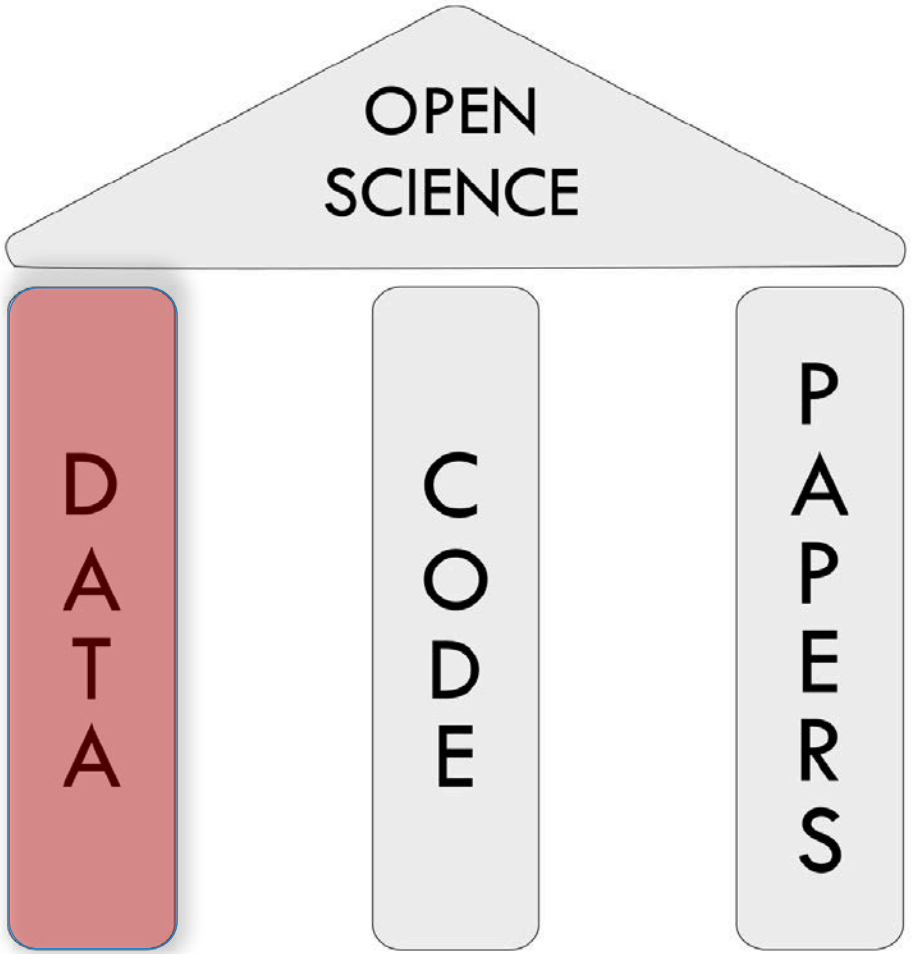


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

Data deposited in a public, community-recognized repository with a stable DOI

Follows FAIR Principle

- Findable
- Accessible
- Intra-operable
- Reusable

Should be deposited *before* publication

Open Data: Community recognized Repositories

Domain Specific Repos

- OpenfMRI
- dbGap
- FCP/INDI
- LONI
- NITRC
- XNAT Central

Data Agnostic Repos

- FigShare
- Dryad
- DataVerse
- Open Science Framework
- NIMH Data Archive (NDAR)

What is Open Code?

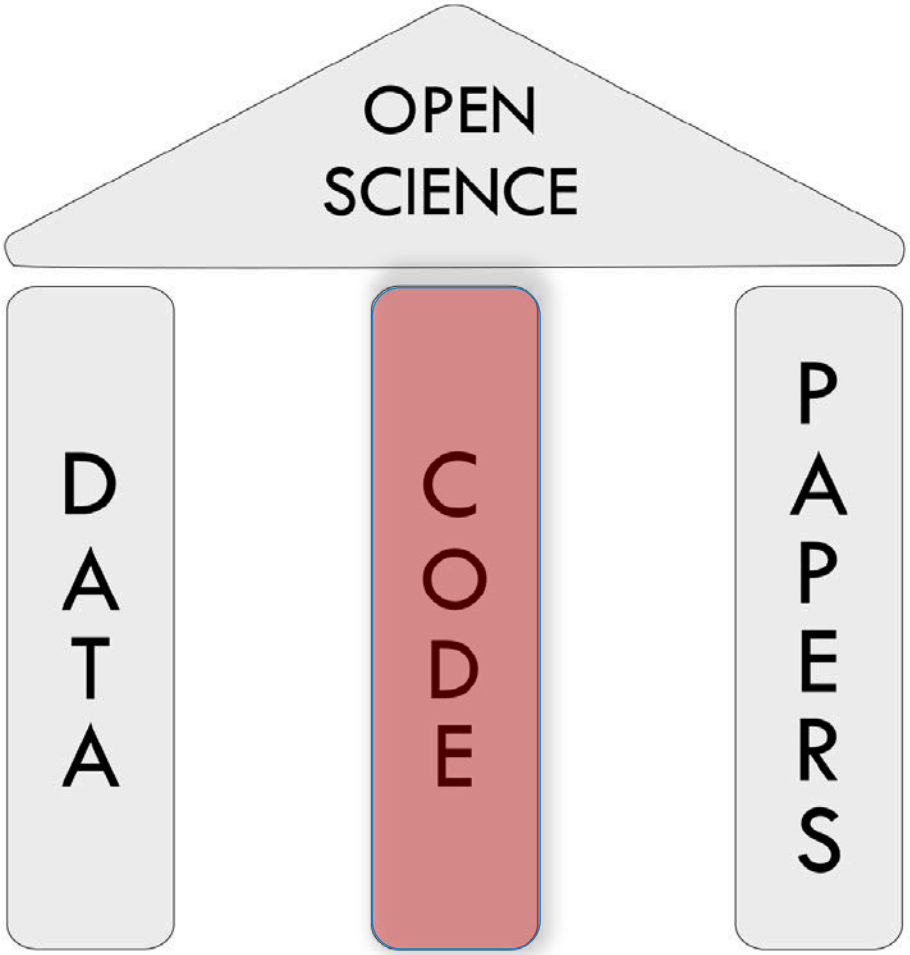


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

Open code enables greater reproducibility (includes non-code methods)

Open Code – Don't Reinvent

Reuse and improve

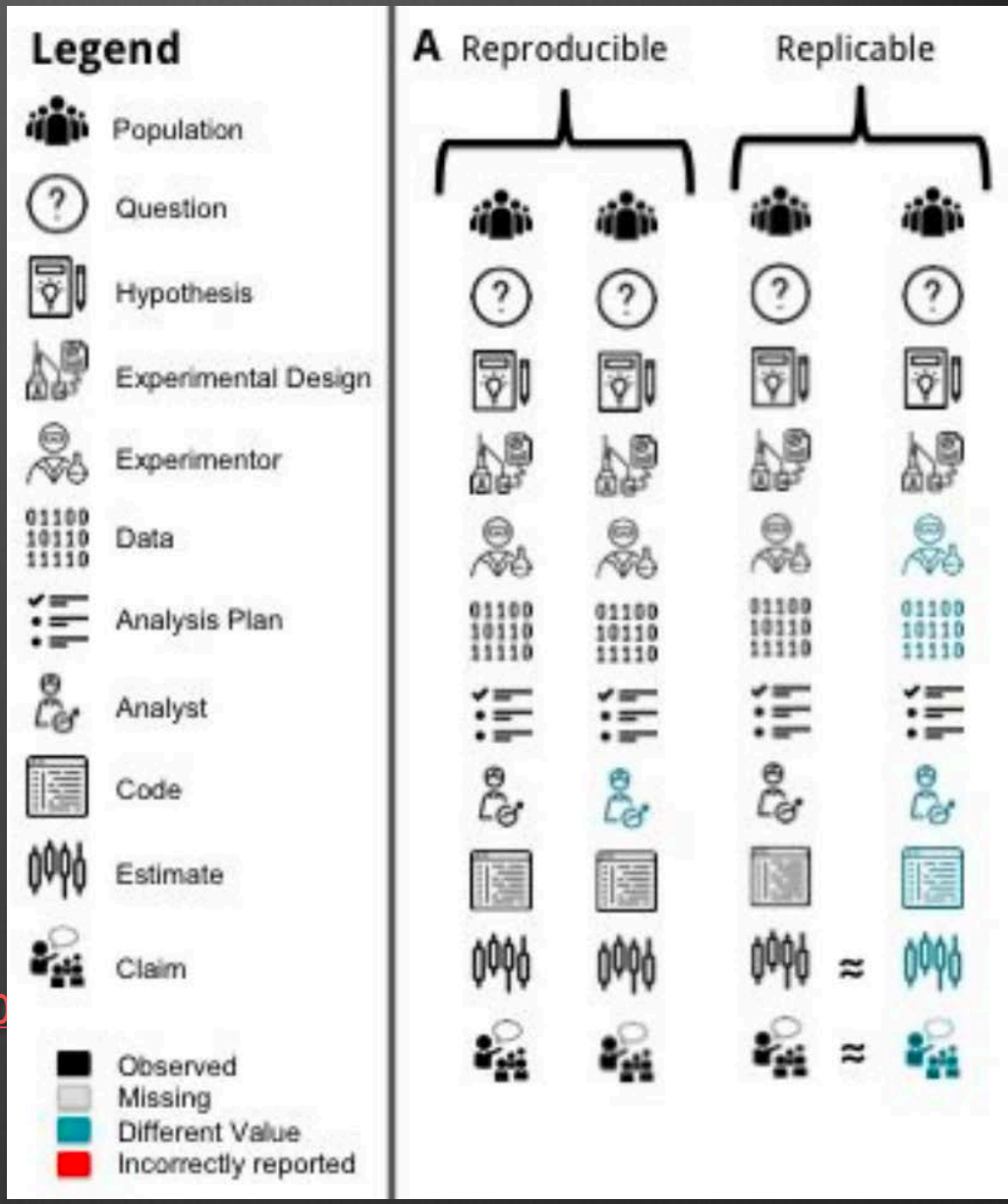
Open Code - Version Control

Version control systems allows you to:

- Store all of your analysis in a central repository
- Keep a history of "snapshots" of your evolving analysis
- Quickly switch between different versions of your analysis
- Adopt and modify code from other scientists
- Collaborate

Why: Replicable vs. Reproducible

- Replication: new data, different team
- Reproducible: Same data, different team
- Where replication is hard or impossible, reproducible is the next best thing
- http://biorxiv.org/content/biorxiv/early/2016/07/29/066803.full.pdf



This Course: Learning Git

- Sticky Notes
- Philosophy: minimize lecturing, mostly interactive and live
- Ask and help your neighbors
- Google Docs for sharing code snippets and links:
  - http://bit.ly/NIHPIDAY2017

What are Open Papers?

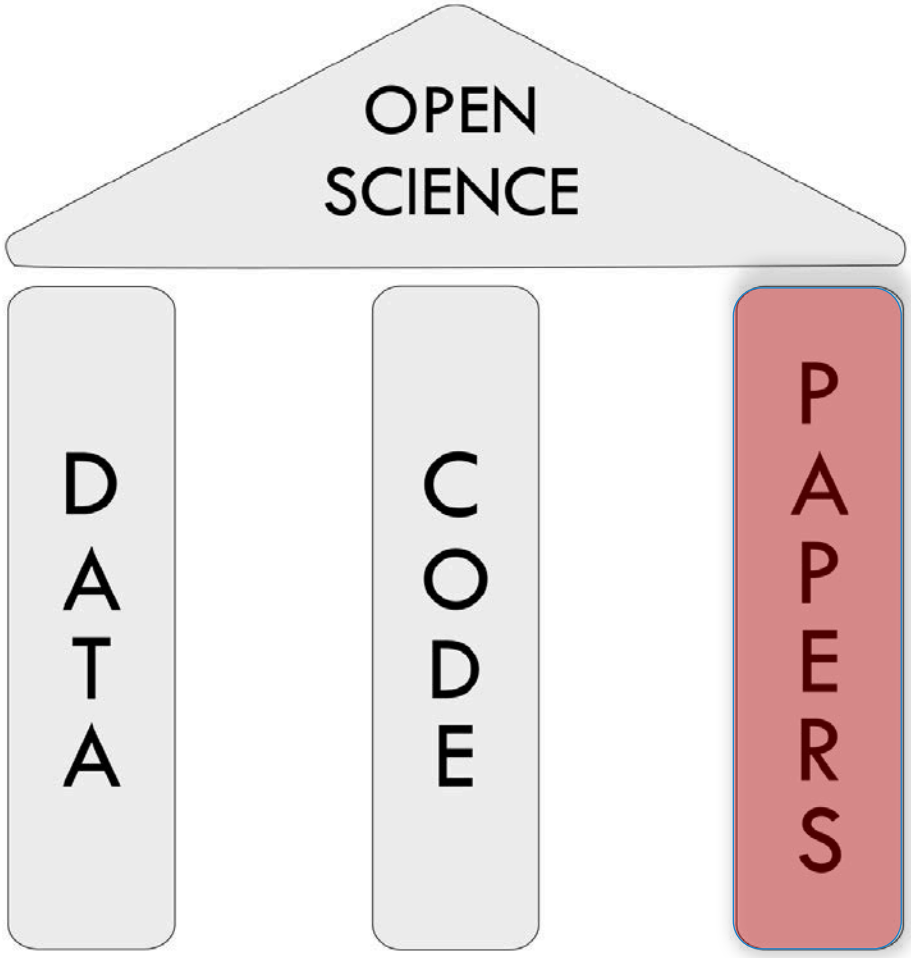


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

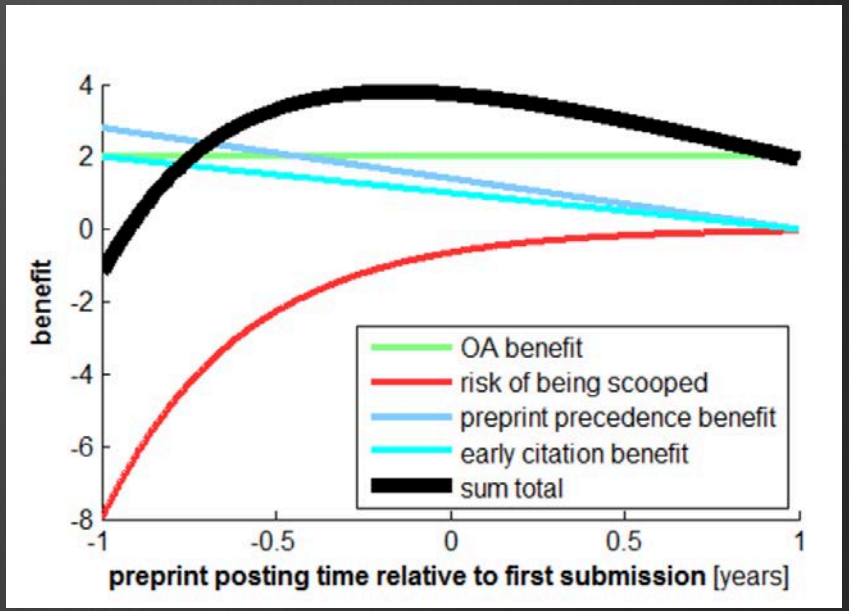
- Preprint posting
- Open access
- Open review

Open Papers: Preprint posting

arXiv.org

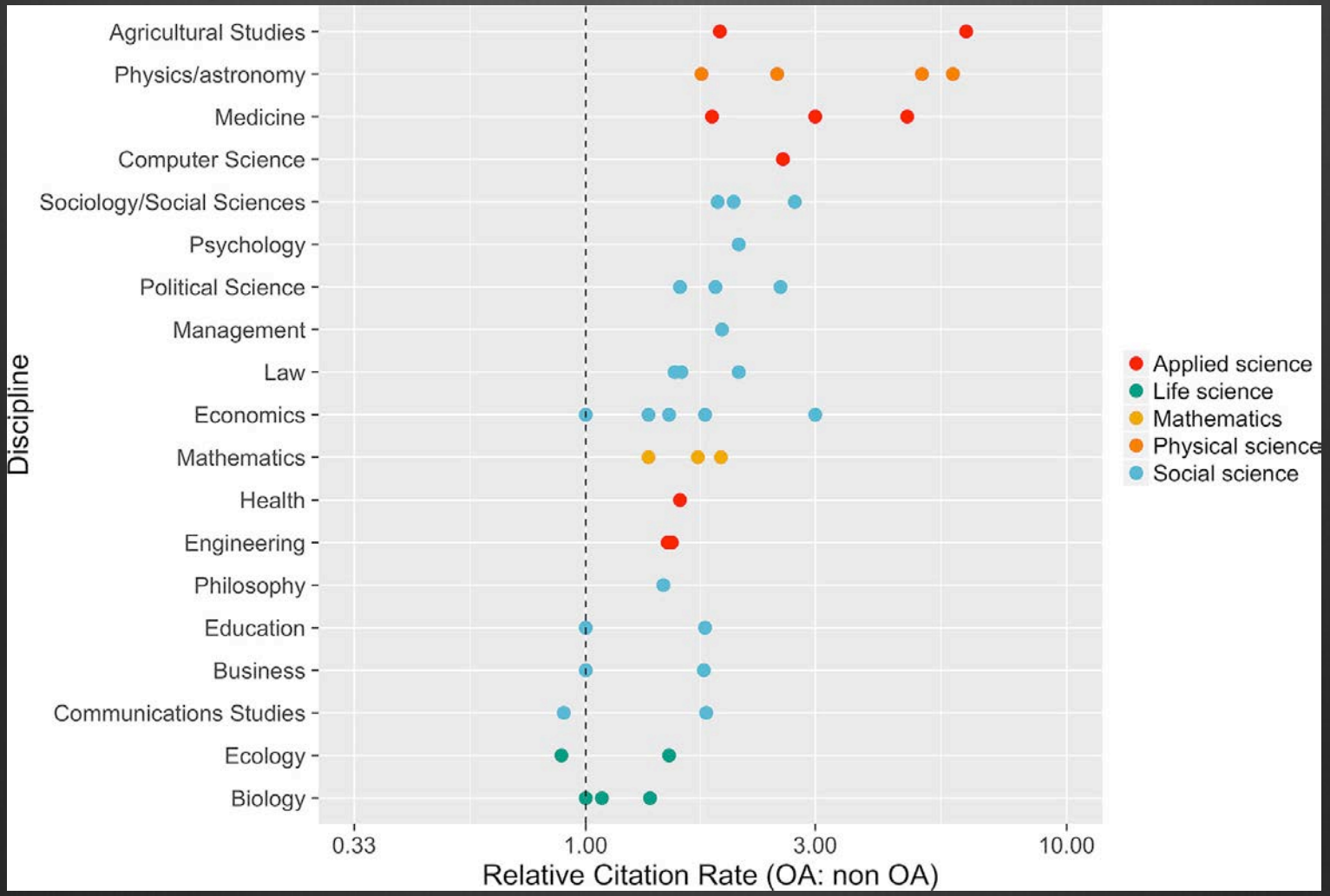
bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

- Benefits:
  - Open access
  - Catch errors
  - Earlier citation
  - Earlier precedence, prevent scooping
  - Speed and improve final submission



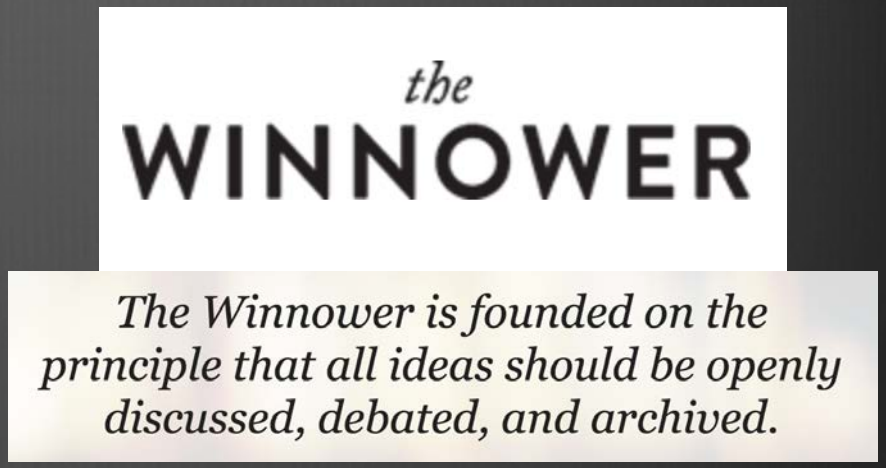
Kriegeskorte (2016) DOI:10.15200/winn.145838.88372

Open Access
Open access publication are cited more



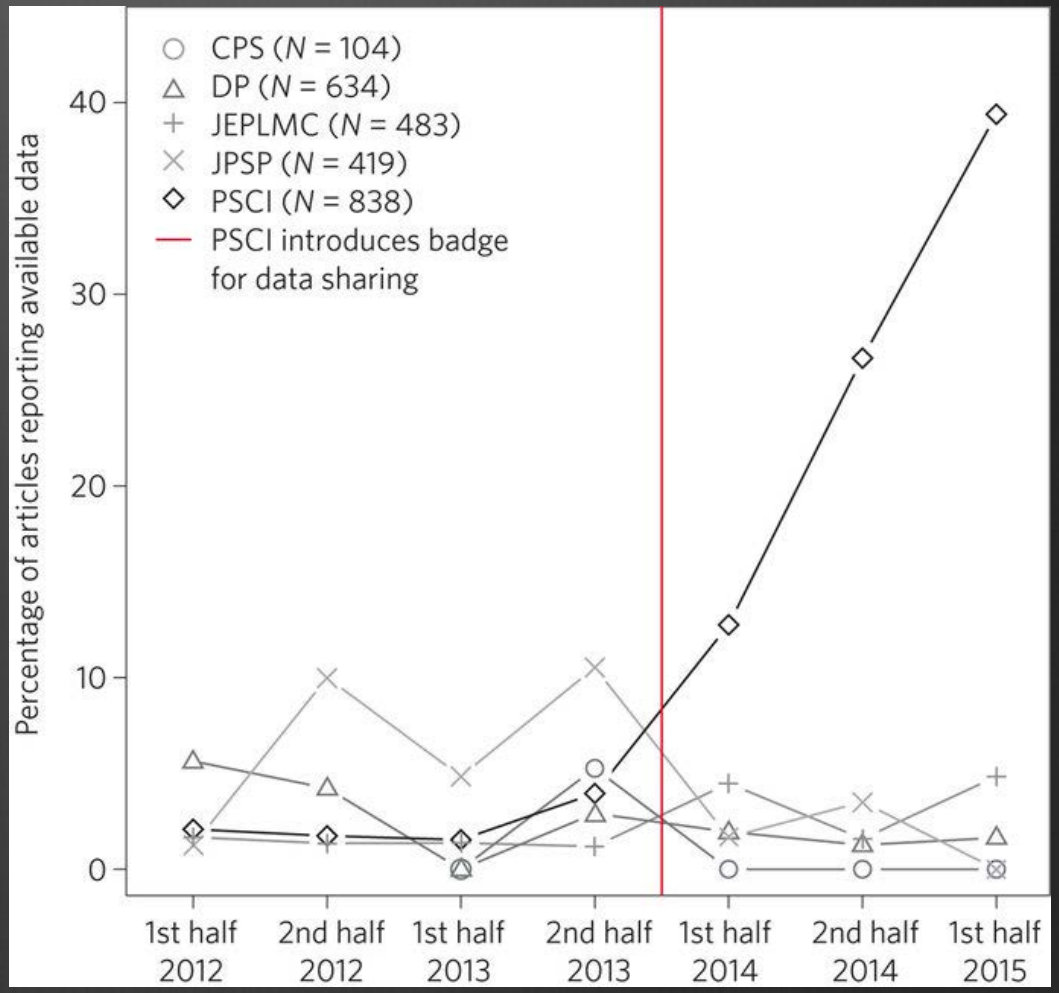
mean citation rate of OA articles divided by mean citation rate of non-OA articles

Open Review



- Public discussion of pros and cons of submission
- Optional anonymity
- Prevent low-quality and or biased review

Incentives: Badges

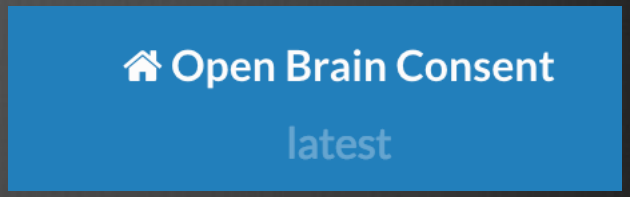


Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

How – Plan Ahead

- Get data sharing in your protocol:
  - NIMH Data Sharing Commitee
  - https://open-brain-consent.readthedocs.io



- When designing, collecting, and analyzing consult with standards documents:
  - Enhancing Quality and Transparency of Health Research (EQUATOR) http://www.equator-network.org

Standards – EQUATOR & COBIDAS

- EQUATOR: Different standards for different designs
  - RCT, crossover, observational, etc.
- COBIDAS: Recommendation for brain imaging
- Both EQUATOR and COBIDAS focus on reporting,
- Reviewing them in advance will help you plan and design your study
- Also useful reference when reviewing papers

Standards – EQUATOR

Checklists

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as p	
	3b	Important changes to methods after	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the dat	
Interventions	5	The interventions for each group wit actually administered	
Outcomes	6a	Completely defined pre-specified pri were assessed	
	6b	Any changes to trial outcomes after	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any	
Randomisation:			

Table D.1. Experimental Design Reporting

Aspect	Notes	Mandatory
Number of subjects	*Elaborate each by group if have more than one group.*	
Subjects approached		N
Subjects consented		N
Subjects refused to participate	Provide reasons.	N
Subjects excluded	Subjects excluded after consenting but before data acquisition; provide reasons.	N
Subjects participated and analyzed	Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.	Y
Inclusion criteria and descriptive statistics	*Elaborate each by group if have more than one group.*	
Age	Mean, standard deviation and range.	Y
Sex	Absolute counts or relative frequencies.	Y
Race & ethnicity	Per guidelines of NIH or other relevant agency.	N

How to be Open – Choose your battles

Be open when you can, as you can

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

How to Open – You don't have to do it alone

- Training

- Asking for help
  - HPC Walk-in sessions
  - NIMH Data Science and Sharing Team

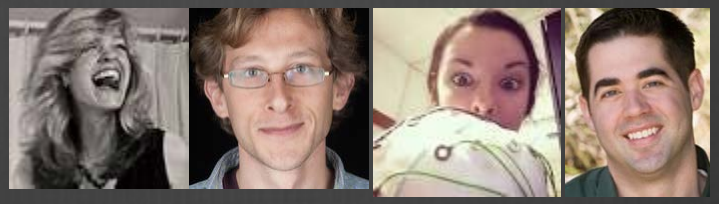
Adam Thomas

John Lee

Dylan Nielson

Data Science and Sharing Team's Workshop on Open and Reproducible Neuroscience Mar 13-17th, 2017

- 45 applications, 25 students attended
- 16 hours of instruction on Python, Git, Data Repositories, Biowulf integration, Pre-registration, and statistical rigor
- Instructors from Gallaudet, King's College London, AFNI and Biowulf Teams



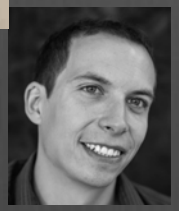
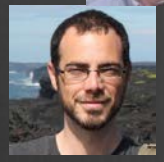
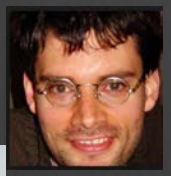
- All course material available online: https://github.com/nih-fmrif/NIMH_repro_2017
- Next course Nov 2017

Summary and Take Homes

- Science is changing (for the better) in both scope (big) and culture (open) to address future challenges
- Open science strives to maximize reproducibility and transparency of data, code, and papers
- Adopting Open Science practices yields benefits in productivity, impact, and reach
- You don't have to do it all at once, and you don't have to do it alone

Credits

Material borrowed, adapted, and/or stolen from:

- Russ Poldrack 
- Chris Gorgolewski 
- Brian Nosek 
- Tal Yorkoni 
- Niko Kriegeskorte 
- Tom Nichols 
- Phil Bourne 

Thanks!

See online slides for more URLs and references:
https://github.com/agt24

Questions?